Mini Project
M1 BDDS

Nadine Laabidi **Khouloud Chalbi**

I. System Architecture

1. Why do we use Big Data technologies in e-health systems ?

The healthcare industries yield the most critical and massive amount of data from various sources such as biomedical research, hospital records, clinical records of patients, clinical examination results, and different health IoT devices.So data collection and analysis enables doctors and health administrators to make more informed decisions about treatment and services and to capture a comprehensive picture of patient experience.

- Volume

The volume of Health Data collected from clinical tests,lab tests, physician visits, administrative data surrounding payments and payers is Huge and it is already expanding.
.Today, approximately 30% of the world's data volume is being generated by the healthcare industry. By 2025, the compound annual growth rate of data for healthcare will reach 36%.

- Variety

Health data is gathered from numerous sources including electronic health records , medical imaging, genomic sequencing, payor records,smartphones apps, pharmaceutical research, wearables, and medical devices. So the Data collected have different forms.

- Velocity

Every second, an exponential amount of healthcare data is generated and mined for valuable insights.So every second, **real-time** decisions must be taken based on real-time Data collection for the sake of the patients' health.

- Value

Healthcare Data is one of the most valuable data. Healthcare organizations work with health data, to make faster crucial decisions and to develop many needed healthcare improvements that:

- speed up the development of new medical products and treatments for individuals who need them.

- Identify risk factors and speed up diagnosis
- Identify pathways in disease transmission, thus preventing diseases or conditions.
- Predict outcomes and increase the effectiveness of treatments.
- Improve the quality and safety of treatments.
- Enhance public health strategy.
- Improve patient care
- support research organizations and scientific associations to develop new treatments and devices.

- Veracity

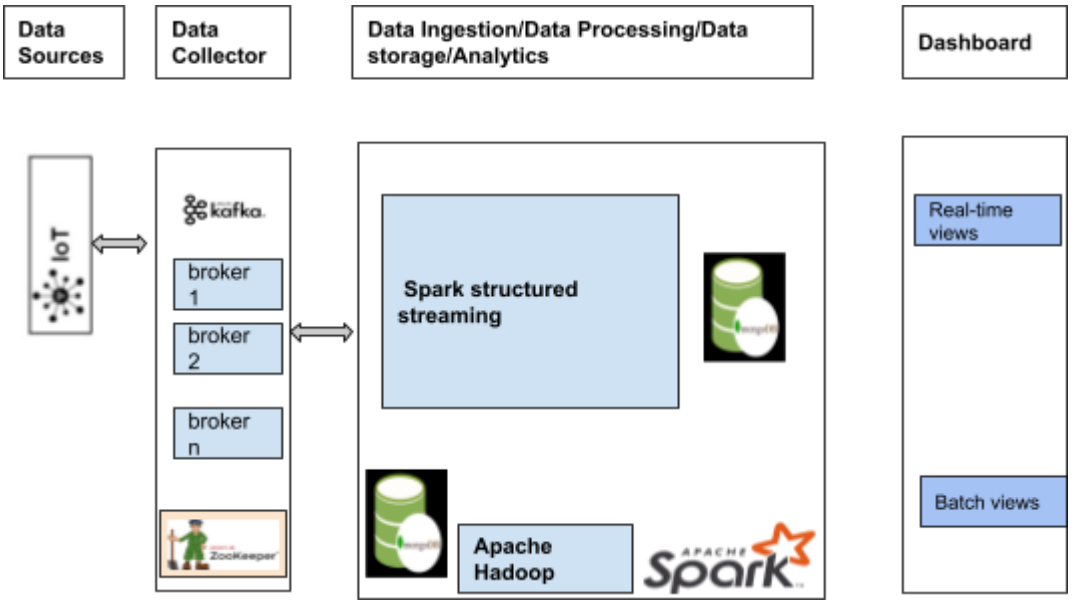
Health Data is accurate, reliable, and trustworthy.

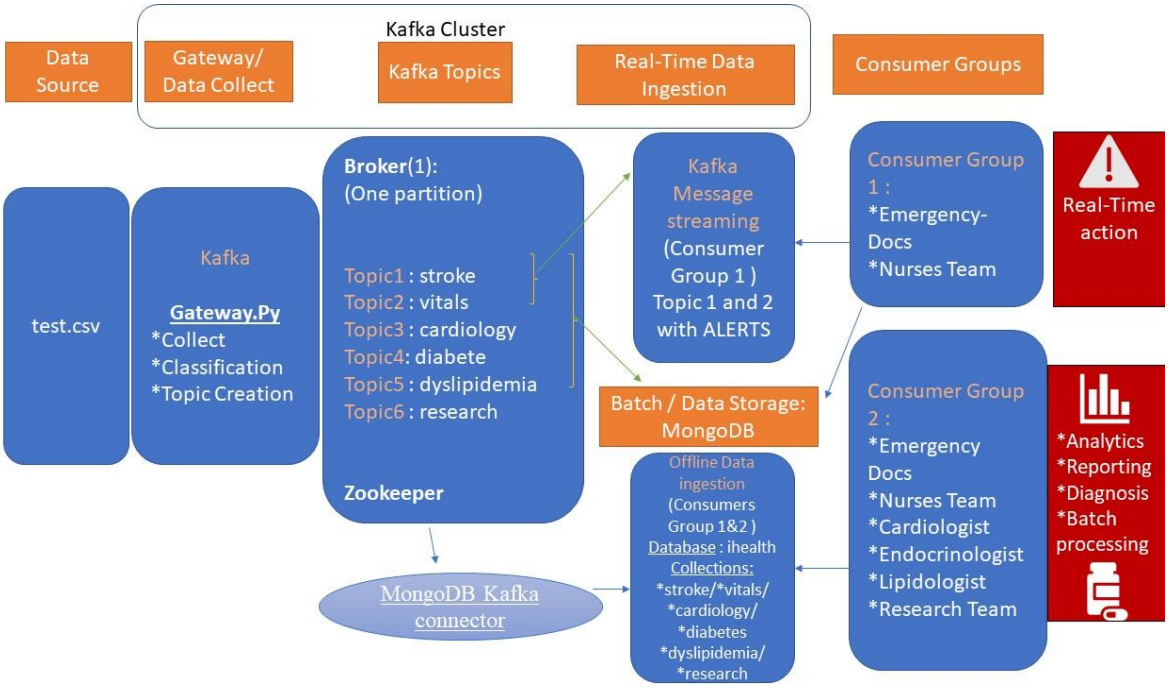
2. Which part of the Data is applied for the stream analysis and which part is for data processing ?

Data that are generated by smart sensors attached to the patients and to their augmented environment is used in real time to provide early warnings of changes in a patient's condition (vital signs).It displays vital parameters such as heart rate, blood-oxygen saturation, respiratory, room temperature…It helps healthcare providers to apply continuous decisions, allowing them to capture and analyze data just in time, speed up the diagnosis and immediately help patients with crucial conditions. This data is processed in real-time : Stream Analysis.

While batch processing is where the processing happens of blocks of data that have already been stored over a period of time.For example, processing the patients data that have been collected in a day/week/year. This collected Data is for researchers, doctors and Data Scientists looking to study the data/develop new treatments/predict diseases/ consult the patients' record…

3. **The reference architecture for this project**





In our case :

- Our producer script in Kafka will do the job of data collector / Gateways and will read / collect data from the csv file line by line to simulate the gateway.

- MongoDb will play the role of Our Storage Sink

- Consumers will be split into 2 groups : Real-time viewers and Batch viewers

To keep it simple, we have

- Kafka cluster (with a Broker and a zookeeper)
- Producer API: It publishes messages to the topics in the Kafka cluster.
- Consumer API:It Consumes messages from the topics in the Kafka cluster.
- Connect API: Directly connect the Kafka cluster to the sink system. The system here is our mongoDB : aNoSQL database.

- **Data sources**

The Data Source Layer is the layer where the data from the source is encountered and subsequently sent to the Data collector. In healthcare systems data sources are usually sensors, software and smart applications. In our case,it is a simulation of a Gateway.

- **Data collector**

The Data Collection layer as the name suggests is responsible for connecting to the source systems and bringing data into the data platform in a periodic manner. It is needed to integrate multi-source, structured and unstructured data for further management. It is required to handle multi-source data efficiently, extract information, and integrate diversified data sources.
For this layer, we chose Kafka because it is a publish-subscribe based messaging system designed from the ground up to provide high throughput, fast performance, scalability, and high availability.
it is :

- Open Sourced via Apache
- Free software license
- Fast, low latency system dedicated to high performance
- Scalable, distributed, and robust design
- Runs as a cluster of servers each of which is called a Broker. Each broker can handle large volumes of data
- Highly scalable storage system

In this case, our kafka producer will simulate the collection by reading the data from our dataset.

- **Stream Processing**

Stream processing allows us to process data in real time as they arrive and quickly detect conditions within a small time period from the point of receiving the data. Stream processing allows you to feed data into analytics tools as soon as they get generated and get instant analytics results.

Streaming data processing applications help with live dashboards, real-time online recommendations, and instant fraud detection.

- Batch Processing

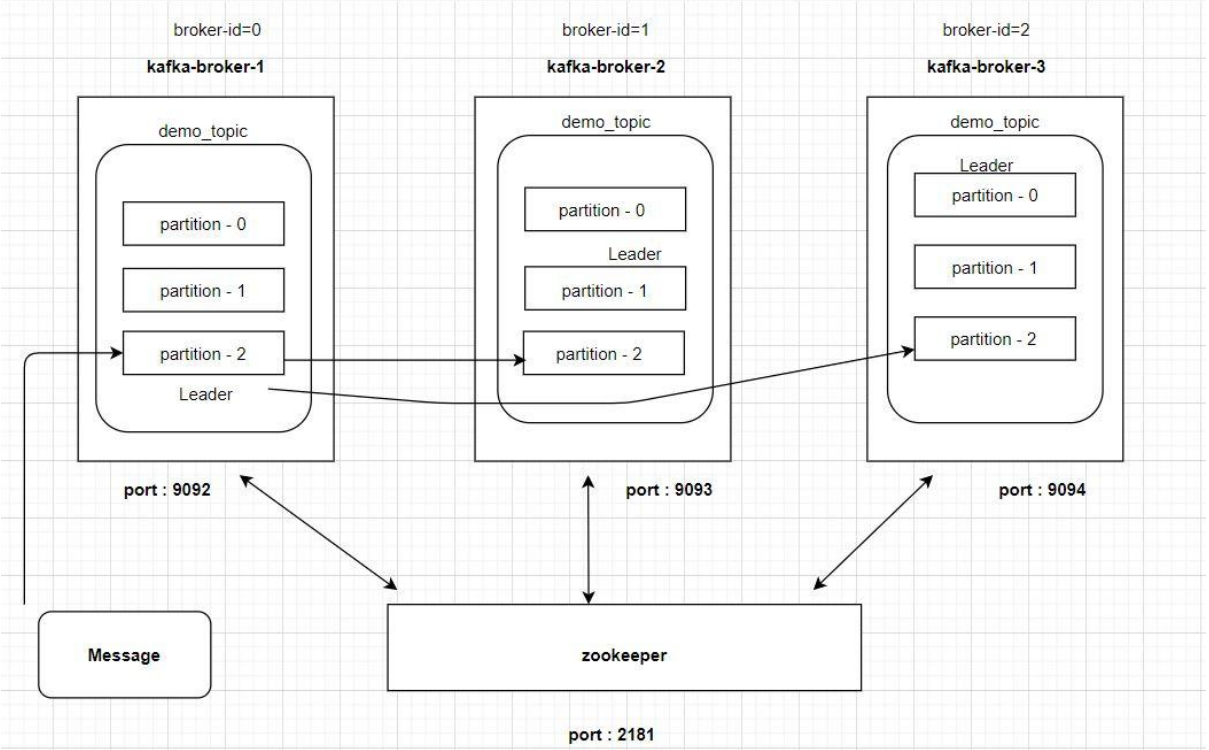
Batch processing processes data quickly and improves the efficiency of job processing. In our case, we used it for creating a new Database to enable healthcare users to access the patients_records / insert new ones/ manage the updates.

- **Batch /Data storage**

This layer of Big Data Architecture focuses on “where to store such large data efficiently.”
It processes data quickly and improves the efficiency of job processing. In our case, we used it for creating a new Database to enable healthcare users to access the patients_records / insert new ones/ manage the updates. We chose MongoDB for its horizontal scaling, powerful query capacity, and document flexibility. We specifically used these features to support various clinical information formats regulated by local legal regulations.

4. Message Broker

A kafka cluster is composed of multiple brokers and each broker contains a certain topic partition. They are responsible for writing new events to partitions, serving reads on existing partitions, and replicating partitions among themselves. So with multiple brokers, the same message is stored multiple times. This **ensures durability as the same message is located on different brokers**. In case of a broker failure, Kafka can switch the leader and provide the replicated message to its clients.



5. Kafka Topics

a. stroke

A stroke is a serious life-threatening medical condition that happens when the blood supply to part of the brain is cut off. Strokes are a medical emergency and urgent treatment is essential.
Stroke will be defined based on specific features values :

- cp
- Trestbps
- restecg

b. vitals

Vital signs are measurements of the body's most basic functions.Theses Vitals need to be constantly monitored. The 3 main vital signs routinely checked by healthcare providers include:

- trestbps - resting blood pressure (in mm Hg on admission to the hospital)
- restecg - resting electrocardiographic results
- tmp -Body temperature

c. cardiology

Angina is chest pain caused by reduced blood flow to the heart muscles. It's not usually life threatening, but it's a warning sign that you could be at risk of a heart attack or stroke.
Classification based on XGBoost : normal patient / Cardio patient

d. diabete

Diabetes is a chronic (long-lasting) health condition that affects how your body turns food into energy and it is detected by a Fasting blood sugar test.

e. dyslipidimea

It is the imbalance of lipids such as cholesterol, low-density lipoprotein cholesterol, (LDL-C), triglycerides, and high-density lipoprotein (HDL).
This topic takes as features : Cholesterol level

f. Research

Healthcare Data is one of the most important data and organizations working with health data are developing many needed healthcare improvements for that the department research needs to have access to the full Database collected.

6. Kafka Consumers :

- **Consumer Group A :**

Nurses Team + Emergency doctors : These consumers are Stream Consumers + get alerted when needed
Emergency doctors get an ALERT if a stroke is predicted!!!
The Nurses Team gets an alert if vital signs exceed normal values!

- **Consumer Group B :**

This group of consumers can read the data from mongodb (Batch/offline). Each user has access to specific Data based on their speciality. For that, as administrators we created new users and gave each one of them access to the Database with specific privileges.

It consists of Different speciality doctors:

- (Cardiologist / Endocrinologist / Lipidologist)
- Research Team
- Nurses team and Emergency doctors who can also see data offline.

7. MongoDB Kafka Connector

The MongoDB Kafka connector is a Confluent-verified connector that persists data from Kafka topics as a data sink into MongoDB as well as publishes changes from MongoDB into Kafka topics as a data source.

Apache Kafka is an event streaming and batching solution and MongoDB is the world's most popular modern database built for handling massive volumes of heterogeneous data. Together MongoDB and Kafka make up the heart of many modern data architectures today. Integrating Kafka with external systems like MongoDB is done through the use of Kafka Connect. This API enables users to leverage ready-to-use components that can stream data from external systems into Kafka topics, and stream data from Kafka topics into external systems.

Mongo Db

Management of users that were given access to the database is the sole responsibility of the user or users with the administrator role: "root" Administrators have the following responsibilities:

- Add new users
- Delete users
- Manage user access
- Set user connection privilege
- Edit user permissions
- View existing user permissions

Change user passwords For our system, we have the following users:

Topic	Mongo Collection	Consumer / User
research	patients_record	researcher
cardiology	cardio_patients	cardiologist
stroke	stroke	emergency_doctor
diabete	diabete_patients	endocrinologist
dylipidemia	dyslipidemia_patients	lipidologist
vitals	abnormal_vitals	nurse